FOCUS ARTICLE AND RESPONSES

Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE)

Mark Davies and Robert Fuchs
Brigham Young University and Münster University

In this paper, we provide an overview of the new GloWbE Corpus — the Corpus of Global Web-based English. GloWbE is based on 1.9 billion words in 1.8 million web pages from 20 different English-speaking countries. Approximately 60 percent of the corpus comes from informal blogs, and the rest from a wide range of other genres and text types. Because of its large size, its architecture and interface, the corpus can be used to examine many types of variation among dialects, which might not be possible with other corpora — including variation in lexis, morphology, (medium- and low-frequency) syntactic constructions, variation in meaning, as well as discourse and its relationship to culture.

Keywords: English, Englishes, global, world, dialects, corpus, corpora, varieties of English

1. Introduction

One of the most important challenges facing researchers of World Englishes is the question of where to find raw data from speakers of these dialects. Possible data sources may include collections of newspapers, blogs, emails, SMS texts, transcripts from recorded conversations, or fictional literature. Studies based on each of these approaches are found in English World-Wide during the past four or five years.

Another possibility is to use “structured corpora”. An important set of corpora for the study of World Englishes is the extended Brown family of corpora, which includes the Brown Corpus of 1960s American English (Francis 1964) and other parallel corpora of varieties and time points such as 1990s American English, 1960s and 1990s British English, as well as Australian English, New Zealand

However, the most widely used corpus for research on World Englishes may be the International Corpus of English (ICE) (see Greenbaum 1996). The ICE components are composed of roughly one million words each (600,000 spoken and 400,000 written), and they currently provide data on 13 national varieties of English, including Great Britain, Ireland, Canada, New Zealand, Hong Kong, East Africa, India, Singapore, the Philippines, and Jamaica, as well as the USA, Nigeria, and Sri Lanka (just the written portion for these last three countries).

As noted, the ICE corpora have been very important for our understanding of World Englishes, as measured by the number of studies that have been based on these corpora. Nevertheless, as valuable as these corpora are, one important limitation is the size of the individual components in ICE. The majority of ICE is composed of transcripts of spoken language, which is extremely difficult and time-consuming to collect. Because the individual corpora have just one million words each, their primary usefulness is probably that they provide the possibility of looking at relatively high-frequency syntactic constructions, where even just a million words might yield enough data. On the other hand, they sometimes do not provide enough data for in-depth research on lexical variation, morphological variation, variation with medium- and lower-frequency syntactic constructions, or differences in word meaning between dialects.

Because of the limitations of smaller corpora, some researchers have created their own proprietary, ad-hoc corpora, in order to study phenomena that need to be based on much larger collections of data. For example, Hundt, Hoffmann and Mukherjee (2012) investigated the use of the hypothetical subjunctive (e.g. *as if he {was/were} king*), where a few million words of data would not have been nearly enough. To collect the needed data, they created a corpus of 146 million words of text from South Asian newspapers (and then compared this to newspaper data from the British National Corpus [BNC]), which provided extremely useful and insightful data for this construction. The downside of such proprietary corpora, however, is precisely the fact that they are proprietary. They are created by individual researchers for use on selected topics, but often are not available to a much wider range of researchers of World Englishes.

Recognizing the need to create a very large corpus of World Englishes, which would be available to a wide range of researchers, we recently collected the GloWbE Corpus (Global Web-based English Corpus). This corpus is based on 1.9 billion words of text from 20 different countries, which includes six Inner Circle and 14 Outer Circle countries (on the distinction between Inner and Outer
Expanding horizons in the study of World Englishes with the 1.9 billion word GloWbE

The texts in the corpus consist of informal blogs (about 60 percent of the corpus) and other web-based materials, such as newspapers, magazines, company websites, and so on. As with the other corpora from corpus.byu.edu, GloWbE is freely available to all researchers at <http://corpus2.byu.edu/glowbe>.

In this paper we provide a number of concrete examples of how GloWbE allows researchers to carry out a wide range of studies on lexical, phraseological, morphological, syntactic, and semantic variation among dialects of English, many of which could probably not be studied with other, smaller corpora. Due to limitations of space, we will provide only a very brief discussion of many phenomena that have been studied in much more detail elsewhere. As a result, there are many aspects of these different phenomena that cannot and will not be exhaustively considered in this paper. But hopefully, this overview of the GloWbE Corpus will show how it can be an important part of researchers’ “toolbox” of resources for studying World Englishes, along with other corpora such as ICE and the Brown-inspired corpora.

2. Designing and creating the GloWbE corpus

There were three goals in the creation of GloWbE: size, genre balance (including informal language), and accuracy in terms of identifying the dialect that it is representing. We will consider each of these goals as we discuss the design and creation of the corpus in this section.

In terms of size, the goal in creating GloWbE was to have a corpus that was large enough to permit research on a wide range of phenomena in World Englishes. To this end, there was really only one possible source for the texts, and that was web pages. Virtually all corpora that are larger than about 500 million words in size are based largely (or exclusively) on web pages. For example, this is the approach used for all of the large corpora from <www.sketchengine.co.uk>. However, as useful as the Sketch Engine corpora are, none of them allow for comparisons between different dialects of English.

But we also wanted to ensure that the web pages represented informal language fairly well. Recall that with the ICE corpora, about 60 percent of the total number of words for each country comes from transcriptions of spoken language, in future updates to the corpus, other countries may be added as well. These may include countries like Malta, Cyprus, Cameroon, Burma, Trinidad and Tobago, and the Bahamas — all of which are former British colonies.

1.
and the other 40 percent consists of more formal, written texts. In the creation of GloWbE, we followed roughly the same approach. About 60 percent of the words for each country come from informal blogs, whereas the other 40 percent come from a wide variety of (often) more formal genres and text types.

The first task in creating the corpus was to get the URLs for millions of web pages from the 20 different countries. In order to do so, we ran hundreds of very high frequency 3-grams (three word strings) in the Corpus of Contemporary American English (COCA) against Google — phrases such as {and from the}, {but if it}, {and they are}, etc. Because of the high frequency of the search strings and because Google does not use search engine optimization criteria for phrases like and from the, it ends up listing essentially random URLs, which is precisely what we wanted. We stored these URLs in a database, along with all associated metadata (web site, country, page title, etc.).

In order to achieve a roughly 60/40 mix of informal and somewhat more formal language, we first collected one million URLs from a “general” search in Google, and then another million URLs from Google searches of just blogs. In the general search, however, about 20 percent of these were also blogs (there is no way to exclude them from “general” searches), which results in (roughly) a 60/40 mix overall.

The most challenging part of the corpus creation was ensuring that the web pages were correctly associated with each of the 20 countries in the corpus. To do so, we carried out each of these two sets of searches (“general” and blogs) for each country separately, using Google “Advanced Search,” and limiting by “Region” (as Google calls it) — Canada, Ireland, India, Singapore, etc.

The question, of course, is how well Google has correctly identified web sites by country. For web sites with a top-level country domain (“.LK” = Sri Lanka, “.SG” = Singapore, etc.) this is quite straightforward for Google. But in the case of “.com” and “.org” web sites, for example, it is much more difficult. In these cases, Google relies on several heuristics, including 1) the IP address for the web server, 2) who links to that website, and 3) who visits the website.

For example, imagine a website <http://www.somesite.com>. The IP address suggests that the server is located in or near Singapore, Google has identified that 95 percent of the visitors to this website come from Singapore and also that 93 percent of the links to this website are from other known websites from Singapore (e.g. those ending in “.SG”). As a result, it is fairly safe for Google to assume that

---

2. The spoken part of ICE contains some formal (scripted) data and the written section includes informal writing (e.g. personal letters). In a sense, then, perhaps the best distinction is “medium”, rather than “formality”.

this website is from Singapore, in spite of the “.com” address. This approach may not be perfect, but it is very good. In the year since we created the corpus (for the first six months it was only available for internal testing), and after checking hundreds of websites to see where they are actually located (based for example on “info” pages on the sites), we have yet to find a single website whose country has not been correctly identified by Google.

After creating the list of URLs, we used HTTrack to download the two million web pages, and we then used JusText to remove “boilerplate” material from the web pages — recurring headers, footers, sidebars, and so on. After this, we used the CLAWS 7 tagger to tag the entire corpus. Finally, we imported the texts into a database, where they would use the same architecture and interface as the other corpora from <corpus.byu.edu>.

The end result was a 1.9 billion word corpus from about 1.8 million web pages in 20 different countries, as shown in Table 1:

Note that the United States (US) and Great Britain (GB) have the largest size (both about 386 million words), all six Inner Circle countries (as well as India) have at least 80 million words of text each, and nearly all of the 20 countries have at least 40 million words of text each (Ghana has 39 million and Tanzania 35 million). As a result of the sampling process, all subcorpora constitute representative samples of how these national varieties of English are used in web-based communication.

As has already been mentioned, GloWbE uses the same architecture and interface as the other corpora from <corpus.byu.edu>. One of the strengths of this architecture and interface is that it allows users to carry out useful comparisons of the different sections of the corpus. In other corpora such as the BNC or the COCA (see Davies 2009, 2011), we can compare and contrast different genres, or, as with the Corpus of Historical American English (COHA; see Davies 2012), we can compare different historical periods. In the case of GloWbE, of course, what we are comparing are the different national varieties. Although we will see many other examples throughout this paper, at this point we will give one quick example of such comparisons.

Suppose that we are looking at the construction “[freak] [p*] out” (forms of freak + pronoun + out, e.g. freaked me out). First, by choosing the “Chart” option under “Display” and entering the search query, we can visualize the data in terms of overall frequency (Figure 1). This shows that the construction is much less com-
mon in the South Asian varieties (IN, LK, PK, BD: India, Sri Lanka, Pakistan, Bangladesh) than in the Inner Circle countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Code</th>
<th>Web sites</th>
<th>Web pages</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>US</td>
<td>82,260</td>
<td>275,156</td>
<td>386,809,355</td>
</tr>
<tr>
<td>Canada</td>
<td>CA</td>
<td>33,776</td>
<td>135,692</td>
<td>134,765,381</td>
</tr>
<tr>
<td>Great Britain</td>
<td>GB</td>
<td>64,351</td>
<td>381,841</td>
<td>387,615,074</td>
</tr>
<tr>
<td>Ireland</td>
<td>IE</td>
<td>15,840</td>
<td>102,147</td>
<td>101,029,231</td>
</tr>
<tr>
<td>Australia</td>
<td>AU</td>
<td>28,881</td>
<td>129,244</td>
<td>148,208,169</td>
</tr>
<tr>
<td>New Zealand</td>
<td>NZ</td>
<td>14,053</td>
<td>82,679</td>
<td>81,390,476</td>
</tr>
<tr>
<td>India</td>
<td>IN</td>
<td>18,618</td>
<td>113,765</td>
<td>96,430,888</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>LK</td>
<td>4,208</td>
<td>38,389</td>
<td>46,583,115</td>
</tr>
<tr>
<td>Pakistan</td>
<td>PK</td>
<td>4,955</td>
<td>42,769</td>
<td>51,367,152</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>BD</td>
<td>5,712</td>
<td>45,059</td>
<td>39,658,255</td>
</tr>
<tr>
<td>Singapore</td>
<td>SG</td>
<td>8,339</td>
<td>45,459</td>
<td>42,974,705</td>
</tr>
<tr>
<td>Malaysia</td>
<td>MY</td>
<td>8,966</td>
<td>45,601</td>
<td>42,420,168</td>
</tr>
<tr>
<td>Philippines</td>
<td>PH</td>
<td>10,224</td>
<td>46,342</td>
<td>43,250,093</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>HK</td>
<td>8,740</td>
<td>43,936</td>
<td>40,450,291</td>
</tr>
<tr>
<td>South Africa</td>
<td>ZA</td>
<td>10,308</td>
<td>45,264</td>
<td>45,364,498</td>
</tr>
<tr>
<td>Nigeria</td>
<td>NG</td>
<td>4,516</td>
<td>37,285</td>
<td>42,646,098</td>
</tr>
<tr>
<td>Ghana</td>
<td>GH</td>
<td>3,616</td>
<td>47,351</td>
<td>38,768,231</td>
</tr>
<tr>
<td>Kenya</td>
<td>KE</td>
<td>5,193</td>
<td>45,962</td>
<td>41,069,085</td>
</tr>
<tr>
<td>Tanzania</td>
<td>TZ</td>
<td>4,575</td>
<td>41,356</td>
<td>35,169,042</td>
</tr>
<tr>
<td>Jamaica</td>
<td>JM</td>
<td>3,488</td>
<td>46,748</td>
<td>39,663,666</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>340,619</td>
<td>1,792,045</td>
<td>1,885,632,973</td>
</tr>
</tbody>
</table>

Figure 1. Overall frequency of *freak out*, by country

In addition to seeing overall frequency, it is also always possible to see the frequency of each individual matching form in each country, as shown in Figure 2. This table view again shows the relatively low frequency of this phrasal verb in the South Asian dialects.
Expanding horizons in the study of World Englishes with the 1.9 billion word GloWbE

Figure 2. Frequency of different forms of *freak out*, by country

In addition to these two main types of visualization, we can also directly compare two different countries or sets of countries — for example, what words occur much more in Ireland than in other countries, or in the South Asian dialects compared to Great Britain. We will see several examples of this in the sections that follow.

Finally, note that for any frequency chart (Figure 1) or table (Figure 2) display, users can click on the desired word and/or country to see the “Keyword in Context” (KWIC) entries (Figure 3).

Figure 3. Keyword in Context (KWIC) display

They can also click on a KWIC line to get even more context — up to about a paragraph — as well as a link to the original web page, to see the complete context. In this paper none of the other 50+ figures show the KWIC lines, but it should be understood that they are fully accessible to users of the GloWbE corpus.

In the following sections, then, we will consider many different phenomena that show how GloWbE can be used to look at variation between dialects for a wide range of phenomena — lexis and phraseology (Section 3), morphology (Section 4), syntax (Section 5), semantics (Section 6), and discourse and culture (Section 7). As we mentioned previously, due to limitations of space, we can only dedicate a paragraph or two to each phenomenon. Because of this, the focus in these sections is on the overall range of possibilities for research with the corpus, and not an in-depth focus on new insights for any of the phenomena themselves.

3. Lexical variation

Previous researchers have noted that lexical frequency is very sensitive to corpus size. For example, Baker (2011: 70) notes that in the Brown family of corpora (one million words each), there might be a few hundred very high frequency words
with enough tokens to compare across dialects (e.g. class, miss, black, true, etc.), but such comparisons would be impossible for the vast majority of words. Other comparatively small corpora like ICE would also be similarly limited.

Because GloWbE is nearly two billion words in size, it is large enough to provide comparisons for many more words. For example, there are more than 100,000 distinct lemmas (101,559) that occur more than 100 times each in GloWbE, and nearly 200,000 (197,270) that occur 25 times each. In this section we will provide a few examples of the rich data on lexical variation in varieties of English which GloWbE provides.

We will begin with a fairly trivial example, simply to show that the GloWbE data does match what our intuition tells us should be happening, and then move on to some more interesting data. First, consider fortnight (Figure 4), which is much more frequent in British English (BrE) than in American English (AmE).

![Figure 4. Fortnight](image)

Of course, with GloWbE we can compare the frequency of any word in any country, compared to the other nineteen countries in the corpus. For example, banjaxed (’ruined, screwed up’) is by far the most common in Irish English (Figure 5).

![Figure 5. banjaxed](image)

Turning to Outer Circle countries, we can see the frequency of hand phone (’mobile / cell phone’) in Malaysian English (Figure 6).

![Figure 6. hand phone](image)

There are of course also lexical items that are limited not just to one country, but rather to a particular region. For example, Eve teas* (’public sexual harassment’) is limited primarily to South Asia (IN, LK, PK, BD) (Figure 7).

![Figure 7. Eve teas*](image)
Expanding horizons in the study of World Englishes with the 1.9 billion word GloWbE

Perhaps even more interesting are those words that are limited mainly to non-Inner Circle dialects, such as *equipments* (Figure 8), which is extremely rare in the Inner Circle countries (Figure 8).

In Figures 4–8, we searched for a particular word that we expected to be more frequent in a particular country or group of countries. But one of the real strengths of GloWbE is that it can quickly search through the database to find all words that are more frequent in one country compared with another. For example, via the web-based interface, we can search for *-ies* plural nouns (*-ies.[nn2*] ) in Australian English (AusE; AU in the examples provided) that are not common in other Inner Circle dialects (US, CA, GB, IE, and NZ): United States, Canada, Great Britain, Ireland, and New Zealand, and the resulting list of words would be the one shown in Figure 9.

Note that not all of the results are examples of the AusE -ies diminutive (e.g. *swannies, telemovies, mesenteries*), but the majority are: *vinnies* (‘wine stores’), *fories* (‘firefighters’), *furphies* (‘rumors’), *dummies* (‘toilets’), *eskies* (‘coolers’), *bikies* (‘bikers’), *tradies* (‘tradesmen’), *pollies* (‘politicians’), *schoolies* (‘schoolchildren’ or ‘breaks from school’), *streeties* (‘homeless people’), and *tanties* (‘tantrums’).
Of course we can also look for phrases, and not just individual words. For example, we can compare phrases like \[\text{be different to}\] (much more common in GB, IE, AU, and NZ than in US or CA; Figure 10) and \[\text{be different than}\] (more common in US and CA; Figure 11).

Other examples of differences in phraseology are \[\text{keep in view}\] in South Asian English (especially in the varieties found in India and Pakistan) (Figure 12), and \[\text{discuss about}\] (Figure 13) in the Outer Circle dialects.\(^6\)

4. Morphological variation

GloWbE can also be used to examine morphological variation among the different dialects of English. To take a fairly obvious and perhaps trivial example, we can search for the frequency of \text{had + \{ gotten / got \}}\), as shown in Figure 14, which is

\(^6\) Note that in the GloWbE search syntax, \([j^*] = [\text{ADJ}]\), \([\text{nn}^*] = (\text{common}) [\text{NOUN}]\), \([p^*] = [\text{PRON}]\), and \([\text{vv}^*] = (\text{lexical}) \text{VERB}\).
based on 13,273 tokens. As we can see, the percentage of all tokens that use gotten (as opposed to got) is three to four times as high in the United States and Canada as it is in Great Britain, which has the lowest percentage of gotten of all of the Inner Circle dialects.

Figure 14. had + {gotten / got}

Another case of competing past participles is [have] + {proved/proven}. There are 29,683 tokens in GloWbE, and they show that proven is much higher in the United States and Canada than in the other Inner Circle dialects (Figure 15). GB, on the other hand, is the Inner Circle dialect that most strongly prefers proved.

Figure 15. [have] + {proven / proved}

GloWbE also shows a huge contrast between the US and CA and most of the other dialects with regards to dove versus dived (see Chambers 1998). There are 1,124 tokens of [pronoun] + {dove/dived} (e.g. he dove into the pool, they dived into their homework). Both US and CA use dove in about 77 percent of all cases, whereas GB is the dialect that uses dove the least, at only 18 percent (Figure 16).

Due to limitations of space in this paper, we have provided just a few examples of how GloWbE can be used to look at morphological differences between varieties of English, and of course many other analyses of morphological variation can be carried out using the corpus.
5. Syntactic variation

In this section we will consider a number of different examples of how GloWbE can be used to carry out investigations of dialectal variation in syntax. We will start with some fairly simple examples, and then progress to more detailed phenomena, which relate to previous research on syntactic variation in English.

To begin with, consider the use of *likely* occurring between verbs (e.g. *they will likely have better careers*), which is discussed in Lindquist (2009: 209–271). Based on data from COCA and the BNC, Lindquist (2009) suggests that the construction is more frequent in AmE than in BrE, and this is clearly supported by the 36,703 tokens of the construction in GloWbE. The data can be visualized either by overall frequency (Figure 17) or by individual string (Figure 18).

A somewhat “broader” construction is the “*way* construction”, which has been widely studied from within the Construction Grammar model (see Goldberg 1997). The simple search for “[vv*]” followed within one or two words by “[ap*] way [i*] the [nn*]” will find all cases of strings like *made his way down the corridor*, *worked her way into the conversation*, *fought his way through the crowd*, and...
Expanding horizons in the study of World Englishes with the 1.9 billion word GloWbE so on. Within just a few seconds, users can search the 1.9 billion words to find the frequency of all 18,525 matching strings, and see all matching verbs and their frequency in each dialect (Figure 19).

Figure 19. “way construction”, by verb and by country

A quick glance at Figure 20 suggests that the “way construction” may be less frequent in the four South Asian varieties (IN, LK, PK, and BD), and this is confirmed by looking at the overall frequency in these dialects.

Figure 20. “way construction”, overall by country

One of the most widely studied constructions during the past two decades or so has been the “quotative like” construction. A consistent theme in most of the recent research is that the construction has now clearly spread beyond being just an American phenomenon, and that it is now found in many other dialects of English (e.g. Tagliamonte and D’Arcy 2004 for Canadian English [CA]; Buchstaller 2006 for BrE; Rodriguez-Louro 2013 for AusE; and D’Arcy 2012 for New Zealand English).

The GloWbE corpus provides very interesting data on the distribution of the construction in blogs and other web pages from the different dialects. The 3,114 tokens of the construction in GloWbE ([c*] [p*] [be] like, ‘e.g. and I was like,’) show that while the construction is still the most common in AmE, it is (in stair-step fashion) progressively less frequent in Canada, Great Britain, Ireland, Australia, and New Zealand, and that it is the least frequent in the South Asian dialects (Figure 21).

Figure 21. “quotative like” construction
Of course we can also use the corpus to look at syntactic phenomena where there is a strong prescriptive norm, and see how this plays out in the different dialects. For example, in AmE, try and VERB is quite stigmatized (e.g. he’ll try and talk to her tomorrow) (see Hommerberg and Tottie 2007). GloWbE has 68,557 tokens of try and, and it clearly is less frequent (Figure 22) in the US (and Canada) than in the other Inner Circle dialects.

![Figure 22. try and [VERB]](image)

In all of the cases of syntactic variation discussed to this point, one simple search in GloWbE was able to provide the needed data. In many cases, however, we will want to combine two successive searches in GloWbE to see the relative frequency of two constructions.

For example, consider verbal complementation with stop (see Rudanko 2002: Chapter 4). Figure 23 shows the frequency without from (they stopped him Ø leaving) and a similar chart can be produced for the construction with from as well (they stopped him from leaving). Here the 21,455 tokens show that the construction without from is much less common in American and Canadian English.

![Figure 23. [stop] + [PRON] + [v-ing]](image)

But we can also input the data from these two charts into a spreadsheet to see the percentage of all complements that lack from. Here we see perhaps even more clearly the relative absence of the –from construction in American and Canadian English (Figure 24).

Another example of dialectal variation is the (prescriptively) non-standard singular (SG) instead of the plural (PL) in cases like there’s/are some people next door. Collins (2012) looks at this construction in eight different ICE corpora and suggests that speakers of Outer Circle varieties are much more reluctant to use the non-standard singular form with plural nominal subjects (e.g. there’s (SG) some people (PL) next door), whereas this is not as much of an issue for speakers of the Inner Circle dialects. The data from the 153,916 tokens in GloWbE support this claim. Table 2 shows the number of tokens with the non-standard singular (e.g.
Expanding horizons in the study of World Englishes with the 1.9 billion word GloWbE

Figure 24. \([\text{stop}] + [\text{PRON}] + [\text{v-ing}] \) (percentage without from)

Table 2. SG/PL agreement with existentials (there is/are)

|       | US  | CA  | GB  | IE  | AU  | NZ  | IN  | PK  | BD  | SG  | MY  | PH  | HK  | ZA  | NG  | GH  | KE  | TZ  | JM  |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SG    | 1018| 377 | 1230| 383 | 582 | 237 | 89  | 23  | 39  | 93  | 58  | 66  | 55  | 58  | 28  | 37  | 35  | 25  | 43  |
| PL    | 29401| 10514| 31220| 7558 | 12432| 6515| 8499| 4100| 3970| 2996| 3405| 3268| 3757| 3052| 3257| 3567| 2866| 3293| 2583| 3155|
| %SG  | 3.3 | 3.5 | 3.8 | 4.8 | 4.5 | 3.5 | 1.0 | 1.0 | 1.1 | 2.7 | 1.7 | 1.7 | 1.8 | 1.7 | 0.8 | 1.3 | 1.1 | 1.0 | 1.3 |
there’s some people next door) and the plural form (e.g. there are some people next door), and the percentage use of the singular by country is shown in Figure 25.

![Figure 25. Percentage SG with existentials (e.g. there’s some people)](image_url)

Table 3 shows the data from another perspective. It depicts the overall percentage of the “non-standard” plural form in all Inner Circle and Outer Circle dialects. These data from GloWbE show that the Inner Circle dialects use the plural form about 2.9 times as frequently as the Outer Circle dialects (3.8/1.3), which agrees very nicely with Figure 2.7 in Collins (2012:67).

Table 3. SG/PL agreement with existentials: Inner versus Outer Circle

<table>
<thead>
<tr>
<th></th>
<th>All Inner Circle</th>
<th>All Outer Circle</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Non-standard&quot; PL</td>
<td>3,827</td>
<td>681</td>
</tr>
<tr>
<td>&quot;Standard&quot; SG</td>
<td>97,640</td>
<td>51,768</td>
</tr>
<tr>
<td>% PL</td>
<td>3.8%</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

A similar case of linguistic conservatism on the part of speakers of Outer Circle varieties deals with another case of verbal agreement — this time with constructions like each (SG) of them (PL) {is|are} — in which the verb can agree with the formal head of the noun phrase (each/SQ) or the notional head (them/PL). Prescriptively, agreement should be with the formal head (each of them is), but many speakers prefer agreement with the notional head (each of them are). To study this construction, we searched for each|none of them|those|these are|were|have for the non-standard plural form and each|none of them|those|these is|was|has for the standard singular form, as shown in Table 4. (Note that these two searches do not find all relevant forms, but those that they do find act as a good “proxy” for other forms, such as each of my friends {is|are}.)

Again, when we compare the Inner and Outer Circle dialects as a group, we see a striking difference. The 7,947 tokens in GloWbE show that speakers of the Inner Circle dialects are more likely to use the innovative, “incorrect” plural than are speakers of the Outer Circle dialects, who are linguistically more conservative.
Table 4. SG/PL agreement with e.g. each of them [is/are]

|     | US  | CA  | GB  | IE  | AU  | NZ  | IN  | LK  | PK  | BD  | SG  | MY  | PH  | HK  | ZA  | NG  | GH  | KE  | TZ  | JM |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| PL  | 1369| 334 | 1046| 237 | 433 | 196 | 208 | 136 | 134 | 93  | 110 | 82  | 68  | 67  | 84  | 62  | 46  | 70  | 47  | 67  |
| SG  | 509 | 188 | 540 | 157 | 215 | 197 | 93  | 62  | 105 | 83  | 74  | 126 | 94  | 101 | 48  | 45  | %PL | 73% | 64% | 66% | 60% | 67% | 66% | 51% | 63% | 46% | 50% | 64% | 44% | 45% | 48% | 52% | 33% | 33% | 41% | 49% | 60% |

Table 6. Subjunctive and indicative with hypotheticals, e.g. as if he [was/were]

|     | US  | CA  | GB  | IE  | AU  | NZ  | IN  | LK  | PK  | BD  | SG  | MY  | PH  | HK  | ZA  | NG  | GH  | KE  | TZ  | JM |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SUBJ| 16707| 3681 | 11983| 2410 | 4111 | 1798| 1654 | 911 | 1247 | 567 | 1254 | 1120 | 1301 | 835 | 835 | 1117 | 683 | 714 | 529 | 788 |
| INDIC| 20969| 5963 | 24581| 5522 | 8474 | 3937| 2859 | 1337 | 2172 | 898 | 1972 | 1660 | 1692 | 1118 | 1939 | 2081 | 1328 | 1406 | 938 | 1798 |
| %SUBJ| 44% | 38% | 33% | 30% | 33% | 31% | 37% | 41% | 36% | 39% | 39% | 40% | 43% | 43% | 30% | 35% | 34% | 34% | 36% | 30% |
And note that the chi square calculation shows that this difference is significant at p < .0001, as it is in Table 3 above.

Table 5. SG/PL agreement with e.g. each of them {is/are}: Inner versus Outer Circle

<table>
<thead>
<tr>
<th></th>
<th>Inner Circle</th>
<th>Outer Circle</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Non-standard&quot; PL</td>
<td>3,615</td>
<td>1,274</td>
</tr>
<tr>
<td>&quot;Standard&quot; SG</td>
<td>1,711</td>
<td>1,347</td>
</tr>
<tr>
<td>% PL</td>
<td>68%</td>
<td>49%</td>
</tr>
</tbody>
</table>

A final case of linguistic conservatism on the part of speakers of Outer Circle dialects deals with the use of the subjunctive in hypotheticals, which was the focus of Hundt, Hoffmann and Mukherjee (2012) for South Asian dialects. In their study, they looked at the frequency of the subjunctive (were) and the indicative (was) after as if, as though, even if, and other cases of if, e.g. he acts as if he {was/were} king. They found that speakers of the South Asian dialects used the older and more conservative subjunctive were more than speakers of BrE.

The data from the 146,889 tokens in GloWbE (Table 6) shows the same — to a point. All four South Asian dialects (shaded in Table 6) use the conservative were more than BrE (33 percent), as well as more than the other Inner Circle dialects of IE, AU, and NZ. Interestingly, AmE (shaded and bolded in Table 6) has the highest degree of the conservative were. The 44 percent figure in AmE is higher than any of the South Asian dialects studied in Hundt, Hoffmann and Mukherjee (2012), and the high degree of the subjunctive in American English confirms what Leech et al (2009) say about diachronic developments in the use of hypothetical subjunctives.

This may be due to strong prescriptive pressure for the use of the subjunctive in AmE (see Auer 2006), just as with the prescriptive pressure to avoid try and, discussed above. The data from GloWbE also show how data from a wide range of dialects of English can provide insights that otherwise might not be available with just a handful of dialects.

To finish this section on syntactic variation, we should note that GloWbE can also be used to look at discourse phenomena, such as discourse markers. We will briefly provide evidence for just one such construction. As Brinton (2009) notes, there is both historical and dialectal variation in the use of the two related discourse markers having said that and that said, which are used to refer back to something that has just been said, and then to provide an alternative point of view. The 8,529 tokens of having said that, show that this variant is much more common in BrE than in AmE (Figure 26), whereas the 11,208 tokens of that said show that this is much more common in AmE, and then progressively less frequent in CA, UK, IE, AU, and NZ (Figure 27).
Expanding horizons in the study of World Englishes with the 1.9 billion word GloWbE

Figure 26. Having said that

Figure 27. That said,

The strong preference for *that said* in AmE in GloWbE ties in nicely with historical data, which shows a large increase in *that said*, over the past few decades. For example, data from 81 tokens in the 400 million word COHA (left side of Figure 28) and 954 tokens in the COCA (right side of Figure 28) shows a clear increase over time. The important number is the frequency per million words, which has risen for example in COCA from 0.39 to 3.83 from the early 1990s to the 2010s.

Figure 28. Historical increase in *that said*,

The ability to relate historical data to questions of dialectal variation is also enhanced in the BYU corpus interface. With just one click, users can seamlessly move from the results of one corpus to another (e.g. COCA or BNC for genre variation, and then GloWbE for dialectal variation, and then COHA for historical variation), and thus easily and quickly explore a wide range of variation in English.

6. Variation in meaning

What corpus-based data could provide evidence for differences in meaning between two or more dialects? For example, how would we know that in AmE *cupboard* is restricted primarily to storing items in a kitchen or pantry, whereas in BrE it can also be used for a storage place in other rooms in the house (cf. American *closet*)? Or how would we know that *scheme* is typically used in a negative sense in AmE, but that this is not the case in other varieties of English?
One approach would be to look at concordance lines for the word or phrase in different dialects, and to see whether the surrounding context might indicate differences in meaning. For example, Figure 29 shows a few of the concordance lines from the GB section of the GloWbE corpus.

Figure 29. Concordance lines for [cupboard]

Notice that in these sentences, the cupboard is over the chimney (#1) or under the stairs (#10), that boxes of photos (#5) or stationary (#7) are stored there, and that it is possible to purchase a stand-alone cupboard (#9) — all of which would seem strange in AmE.

However, given a large enough corpus, we can use another approach. Rather than looking at all 8,726 tokens of the lemma cupboard in GloWbE, for example, we can simply have the corpus interface look for all collocates of cupboard. We can then compare the collocates to see which ones occur in one dialect but not another, and which may therefore signal differences in meaning and usage.

For example, Figure 30 shows a comparison of the collocates of cupboard in 386 million words of AmE (left) and 387 million words from GB (right) in GloWbE.

Figure 30. Collocates of [cupboard]: US (left), GB (right)

While not all of the collocates are of course relevant, many are. For example, refrigerator and pantry are more frequent (per million words) in AmE, probably because there are more references to cupboard in the context of a kitchen. In BrE, on the other hand, there are references to brooms and wardrobes, as well as to skeletons in the cupboard, all of which would be used with closet in AmE. (It is also important to recall that we can click on any collocate to see it in context with cupboard in the corpus, as shown in Figure 3 above.)
Further, let us consider the collocates of scheme in AmE and BrE, as shown in Figure 31. In AmE (left), there are references to alleged, evil, fraudulent, Ponzi, (get) rich quick, and illegal schemes, whereas in BrE (right) the collocates are much more prosaic and neutral in tone (or even positive: note generous, innovative, competent, and qualified). In corpus linguistic terms, we could say that scheme has “negative prosody” in AmE (cf. Louw 1993), whereas this is not the case for BrE.

Figure 31. Collocates of [scheme]: US (left), GB (right)

In these three cases, we compared BrE and AmE. This was done for two reasons. First, these are the two varieties with a global reach, and many speakers of other varieties are familiar with them. Second, these are the two largest segments of GloWbE, at about 385 million words each. Such comparisons may still be possible with smaller segments, perhaps even with countries like Tanzania (35 million words), Ghana (39 million words), or Bangladesh (40 million words), which are among the smallest in the corpus. This is especially the case if regional dialects are compared (e.g. Africa = 203 million words, or South Asia = 234 million words).

7. Discourse and culture

In this section, we will consider “discourse”, in the sense of “which topics of discussion are more common” in one dialect (or groups of dialects) than another, and “what is being said” about particular concepts in different dialects.

At the most basic level, we can use GloWbE to compare the frequency of particular words in different dialects (cf. Section 3) and then consider how this may relate to the culture of those speakers. For example, words with Buddh* (e.g. Buddhist, Buddhism, Buddha) are the most common in Sri Lanka (LK) — the one country in the corpus that is predominantly Buddhist (Figure 32).

Figure 32. Buddh*
Not surprisingly, *Quran* and *Allah* are likewise the most common in Pakistan, Bangladesh, and Malaysia — the three countries in the corpus with the greatest proportion of Muslims. Or consider *feminism*, which is the most common overall in the Inner Circle countries (Figure 33), although the frequency in Ireland (perhaps the most culturally conservative of these countries) is the lowest of these Inner Circle countries.

In addition to simply looking at word frequency, we can also compare the collocates of a given word, to see “what is being said” about particular concepts in different countries. For example, Figure 34 shows the most frequent adjectival collocates of *belief* in South Asia (left) and the six Inner Circle countries (right). Notice the use of *Hindu*, *Muslim*, *Islamic*, *polytheistic*, *sectarian*, and *heretical* in South Asia (all of which are probably related primarily to religion), compared to *liberal*, *deepest*, *positive*, *economic*, *confident*, *causal*, and *non-religious* in the Inner Circle countries (more secular).

Another example of the ability to gain cultural insight from the comparison of collocates are the adjectival collocates of the lemma *marriage* in the Outer Circle countries (left) and the Inner Circle countries (right) in Figure 35. In the Outer Circle countries, there is concern about *inter-caste*, *fixed*, and *forceful* marriages, as well as *permanent* versus *temporary* marriages (perhaps as a husband is forced to look for work outside of his home country). In the Inner Circle countries, on the other hand, people are apparently more concerned with the “hot button” topic of same-sex marriage, with adjectives like *opposite-sex* and *same-sex*, and related words like *anti-gay*, *supporting* and *preserving* (i.e. traditional heterosexual mar-

---

**Figure 33. feminism**

In addition to simply looking at word frequency, we can also compare the collocates of a given word, to see “what is being said” about particular concepts in different countries. For example, Figure 34 shows the most frequent adjectival collocates of *belief* in South Asia (left) and the six Inner Circle countries (right). Notice the use of *Hindu*, *Muslim*, *Islamic*, *polytheistic*, *sectarian*, and *heretical* in South Asia (all of which are probably related primarily to religion), compared to *liberal*, *deepest*, *positive*, *economic*, *confident*, *causal*, and *non-religious* in the Inner Circle countries (more secular).

**Figure 34. Collocates of [belief]: South Asia (left), Inner Circle (right)**

Another example of the ability to gain cultural insight from the comparison of collocates are the adjectival collocates of the lemma *marriage* in the Outer Circle countries (left) and the Inner Circle countries (right) in Figure 35. In the Outer Circle countries, there is concern about *inter-caste*, *fixed*, and *forceful* marriages, as well as *permanent* versus *temporary* marriages (perhaps as a husband is forced to look for work outside of his home country). In the Inner Circle countries, on the other hand, people are apparently more concerned with the “hot button” topic of same-sex marriage, with adjectives like *opposite-sex* and *same-sex*, and related words like *anti-gay*, *supporting* and *preserving* (i.e. traditional heterosexual mar-

---

**Figure 35. Collocates of marriage**
riage), as well as pro-abortion and unborn — apparently referring to “conservatives” and “liberals”, in the context of their views on same-sex marriages.

Figure 35. Collocates of families: Outer Core (left), Inner Circle (right)

Before leaving this section, which deals with how the GloWbE data can provide cultural insights, we want to come back to something that we discussed in Section 3, where we dealt with lexical differences between the dialects. Recall that in that section, we looked for *ics words in Australia, compared to the other Inner Circle countries, and found examples like firies, bikies, and tradies. These are interesting from a lexical point of view, but they provided some insight into cultural differences between the different countries.

We can do similar searches, however, which provide more culturally interesting data. For example, if we compare all *ism words in Great Britain and South Asia, we find the following (Figure 36). In Great Britain (left), people are writing about Euroscepticism, Labourism, presenteeism, nimbysm (nimby = ‘not in my backyard’), monetarism, Thatcherism, and Blairism — with most of these being political in nature. In South Asia, on the other hand, the *ism words are much more related to religion — Qadianism, castism, Talibanism, Vaisnaism, Shivaism, Shiaism, and so on (with the exception of Naxalism). Thus there seems to be a real difference in terms of what people in these two regions are writing about on the Web.

Figure 36. *ism words: Great Britain (left), South Asia (right)

Finally, rather than comparing two countries or sets of countries, we can simply leave the query “open” as far as country goes, and look at all *ism words in all countries. Here we find that discussions of tourism are more common in Africa and Jamaica, people in South Asia are writing more about terrorism, autism is
discussed most in the Inner Circle countries, and the topic of feminism is more common in the Inner Circle countries as well, as we discussed earlier in this section.

8. Conclusion

In Sections 3–7, we have seen a number of examples showing how the data from GloWbE can be used to insightfully investigate a wide range of phenomena in different dialects of English. One aspect of this that we have alluded to throughout the paper, but which we might touch on in a somewhat more detailed fashion here, is the importance of corpus size.

Other than GloWbE, the only other corpus of English that contains data from a number of different dialects, and which is organized in a way that allows us to compare across these dialects, is ICE. As we have discussed previously, ICE contains one million words each for 14 different dialects (11 of which contain both spoken and written English), for a total of about 12,200,000 words of text. GloWbE, on the other hand, contains about 1.9 billion words of data. In other words, GloWbE is more than 150 times larger than ICE. Where ICE may yield 20–30 tokens of a given word, phrase, or construction, GloWbE will often yield 150 times as many, or in other words 3,000–4,000 tokens for the same phenomenon.

Another advantage of GloWbE is that it provides data on a number of varieties so far not included in ICE (such as Pakistani and Malaysian English).

For high frequency syntactic constructions, ICE often has enough data, and this is why it is probably no surprise that so many ICE-based studies in fact deal with rather high frequency constructions. But for many of the phenomena discussed in this paper, ICE would probably not have enough tokens. For example, most of the words and phrases shown in Section 3 occur 500–2000 times in GloWbE, yet they would only occur between perhaps four and 15 times in ICE. In terms of morphological variation, contrasting forms like dived/dove occur
1,000–1,200 times in GloWbE, and they might therefore only occur six or seven times in ICE. In GloWbE there are about 8,000 tokens for a construction like each of them {is|are}, and in ICE there would be only about 50 tokens — probably too few to say much of interest. And things are even more problematic in terms of the number of tokens for collocates shown in Section 6 and Section 7. For a given collocate, there are often only 30–40 tokens in GloWbE, and with a corpus only 1/150th the size, we might be lucky to have a single token in ICE.

But of course size is not everything. The ICE corpora have been constructed very carefully, and for phenomena where “every token counts” and when there can be no “messiness” at all in the data, the carefully-curated, manually annotated ICE corpora may be more useful than GloWbE. Likewise, for phenomena where actual spoken material is needed, ICE will probably be better than GloWbE, where there is no spoken data (although the 60 percent or so of texts in GloWbE that come from blogs do provide fairly informal language). Finally, in GloWbE we only know that a website is from a particular country, but there might be speakers from other countries who have posted to that website. In ICE, on the other hand, care has been taken to ensure that all speakers are from the country in question.

In other words, it is probably not an “either/or” issue when it comes to the use of different corpora, in which the use of one corpus precludes the use of another. Researchers may want to use ICE for some studies, GloWbE for others, and perhaps proprietary corpora that they have created for yet other studies. All of these can be seen as useful “tools” in the researchers’ “toolbox”, and they complement each other nicely.

To the extent, though, that researchers do adopt GloWbE as part of their “toolbox” (along with ICE and other corpora), they will be able to expand their horizons in terms of the types of variation that they consider, as they carry out research on World Englishes.

References


Rodríguez Louro, Celeste. 2013. "Quotatives Down Under: Be like in Cross-Generational Australian English Speech". *English World-Wide* 34: 48–76. DOI: 10.1075/eww.34.1.03rod


**Author’s address**

Mark Davies  
Dept. Linguistics and English Language  
4071 JFSB  
Brigham Young University  
Provo, UT 84602 USA